Assignment 10: Data Scraping

Laila Abed

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on data scraping.

Directions

1. Rename this file `<FirstLast>_A10_DataScraping.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

- Load the packages `tidyverse`, `rvest`, and any others you end up using.
- Check your working directory

```
#1
library(tidyverse)
library(rvest)
```

2. We will be scraping data from the NC DEQs Local Water Supply Planning website, specifically the Durham's 2023 Municipal Local Water Supply Plan (LWSP):

- Navigate to https://www.ncwater.org/WUDC/app/LWSP/search.php
- Scroll down and select the LWSP link next to Durham Municipality.
- Note the web address: https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023

Indicate this website as the as the URL to be scraped. (In other words, read the contents into an `rvest` webpage object.)

```
#2
webpage <- read_html(
  'https://www.ncwater.org/WUDC/app/LWSP/report.php?pwsid=03-32-010&year=2023')
```

3. The data we want to collect are listed below:

- From the “1. System Information” section:
- Water system name
- PWSID
- Ownership
- From the “3. Water Supply Sources” section:
- Maximum Day Use (MGD) - for each month

In the code chunk below scrape these values, assigning them to four separate variables.

> HINT: The first value should be “Durham”, the second “03-32-010”, the third “Municipality”, and the last should be a vector of 12 numeric values (represented as strings)“.

```
#3

Water_system_name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
PWSID <- webpage %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
Ownership <- webpage %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
Maximum_Day_Use <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
```

4. Convert your scraped data into a dataframe. This dataframe should have a column for each of the 4 variables scraped and a row for the month corresponding to the withdrawal data. Also add a Date column that includes your month and year in data format. (Feel free to add a Year column too, if you wish.)

> TIP: Use `rep()` to repeat a value when creating a dataframe.

> NOTE: It’s likely you won’t be able to scrape the monthly widthrawal data in chronological order. You can overcome this by creating a month column manually assigning values in the order the data are scraped: “Jan”, “May”, “Sept”, “Feb”, etc… Or, you could scrape month values from the web page…

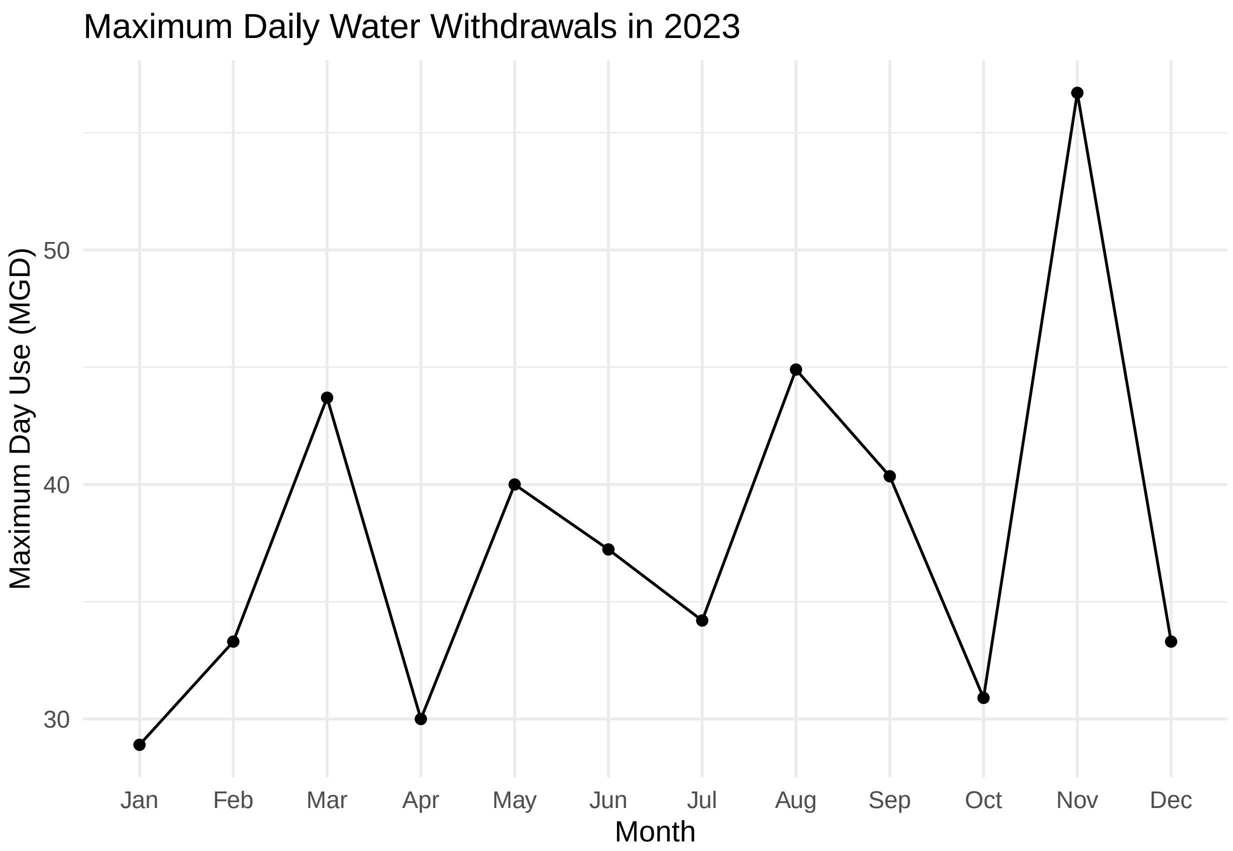
5. Create a line plot of the maximum daily withdrawals across the months for 2023, making sure, the months are presented in proper sequence.

```
#4
#Creating vectors for months and years
months <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug",
              "Sep", "Oct", "Nov", "Dec")
years <- rep(2023, times = 12)
# Matching the length of the months vector
```

```r
water_system_name_vector <- rep(Water_system_name, times = 12)
PWSID_vector <- rep(PWSID, times = 12)
ownership_vector <- rep(Ownership, times = 12)
# Creating a dataframe
water_data <- data.frame(
WaterSystemName = as.vector(Water_system_name),
PWSID = as.vector (PWSID),
Ownership = as.vector (Ownership),
Month = months,
Year = years,
MaxDayUseMGD = as.numeric(Maximum_Day_Use)
)
# Adding a Date column
water_data$Date <- make_date(year = water_data$Year,
                  month = match(water_data$Month, months), day = 1)
# Viewing the dataframe
print(water_data)
```

```
##    WaterSystemName     PWSID    Ownership Month Year MaxDayUseMGD       Date
## 1           Durham 03-32-010 Municipality   Jan 2023        28.90 2023-01-01
## 2           Durham 03-32-010 Municipality   Feb 2023        33.30 2023-02-01
## 3           Durham 03-32-010 Municipality   Mar 2023        43.70 2023-03-01
## 4           Durham 03-32-010 Municipality   Apr 2023        30.00 2023-04-01
## 5           Durham 03-32-010 Municipality   May 2023        40.00 2023-05-01
## 6           Durham 03-32-010 Municipality   Jun 2023        37.23 2023-06-01
## 7           Durham 03-32-010 Municipality   Jul 2023        34.20 2023-07-01
## 8           Durham 03-32-010 Municipality   Aug 2023        44.90 2023-08-01
## 9           Durham 03-32-010 Municipality   Sep 2023        40.35 2023-09-01
## 10          Durham 03-32-010 Municipality   Oct 2023        30.90 2023-10-01
## 11          Durham 03-32-010 Municipality   Nov 2023        56.70 2023-11-01
## 12          Durham 03-32-010 Municipality   Dec 2023        33.30 2023-12-01
```

```r
#5
#Check the Month if it is correct order
water_data$Month <- factor(water_data$Month, levels = c("Jan", "Feb", "Mar",
             "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
# Creating the line plot
ggplot(water_data, aes(x = Month, y = MaxDayUseMGD, group = 1)) +
geom_line() +
geom_point() +
labs(title = "Maximum Daily Water Withdrawals in 2023",
x = "Month",
y = "Maximum Day Use (MGD)") +
theme_minimal()
```



```
print(water_data)
```

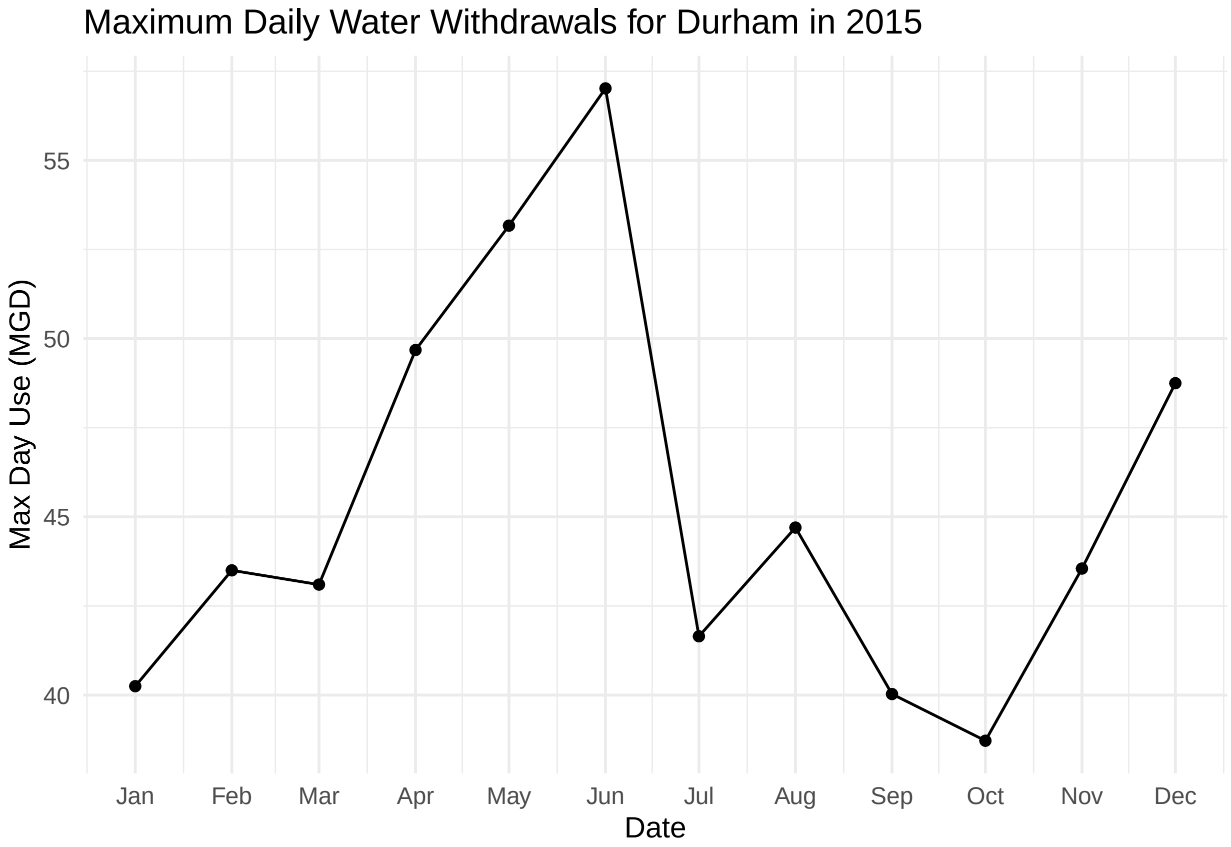
```
##    WaterSystemName     PWSID    Ownership Month Year MaxDayUseMGD       Date
## 1           Durham 03-32-010 Municipality   Jan 2023        28.90 2023-01-01
## 2           Durham 03-32-010 Municipality   Feb 2023        33.30 2023-02-01
## 3           Durham 03-32-010 Municipality   Mar 2023        43.70 2023-03-01
## 4           Durham 03-32-010 Municipality   Apr 2023        30.00 2023-04-01
## 5           Durham 03-32-010 Municipality   May 2023        40.00 2023-05-01
## 6           Durham 03-32-010 Municipality   Jun 2023        37.23 2023-06-01
## 7           Durham 03-32-010 Municipality   Jul 2023        34.20 2023-07-01
## 8           Durham 03-32-010 Municipality   Aug 2023        44.90 2023-08-01
## 9           Durham 03-32-010 Municipality   Sep 2023        40.35 2023-09-01
## 10          Durham 03-32-010 Municipality   Oct 2023        30.90 2023-10-01
## 11          Durham 03-32-010 Municipality   Nov 2023        56.70 2023-11-01
## 12          Durham 03-32-010 Municipality   Dec 2023        33.30 2023-12-01
```

6. Note that the PWSID and the year appear in the web address for the page we scraped. Construct a function using your code above that can scrape data for any PWSID and year for which the NC DEQ has data, returning a dataframe. **Be sure to modify the code to reflect the year and site (pwsid) scraped**.

```
#6.
scrape_data_function <- function(the_PWSID, the_year) {
the_base_url <- 'https://www.ncwater.org/WUDC/app/LWSP'
the_scrape_url <- paste0(the_base_url, '/report.php?pwsid=', the_PWSID,
                  '&year=', the_year)
webpage <- read_html(the_scrape_url)
Water_system_name <- webpage %>%
  html_nodes("div+ table tr:nth-child(1) td:nth-child(2)") %>%
  html_text()
PWSID <- webpage %>%
  html_nodes("td tr:nth-child(1) td:nth-child(5)") %>%
  html_text()
Ownership <- webpage %>%
  html_nodes("div+ table tr:nth-child(2) td:nth-child(4)") %>%
  html_text()
Maximum_Day_Use <- webpage %>%
  html_nodes("th~ td+ td") %>%
  html_text()
#creating the dataframe
df_withdrawals_func <- data.frame("Month" = c(
  1, 5, 9, 2, 6, 10, 3, 7, 11, 4,8,12),
"Year" = rep(the_year,12),
"MaxDayUsage" = as.numeric(Maximum_Day_Use)) %>%
mutate(WaterSystemName= !!Water_system_name,
PWSID=!!PWSID,
OOwnership=!!Ownership,
Date=lubridate::my(paste(Month, "-", Year))) %>%
dplyr::arrange(Date)
#Return the dataframe
return(df_withdrawals_func)
}
```

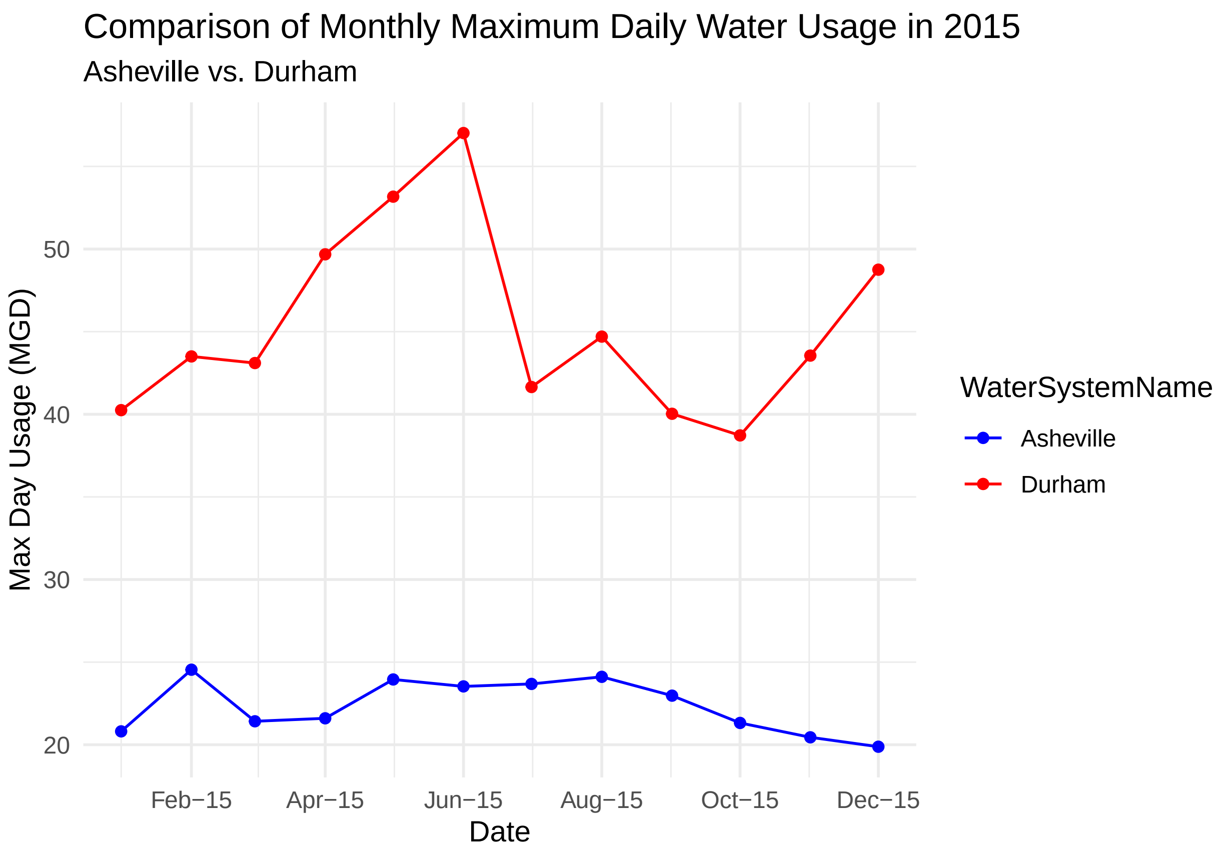
7. Use the function above to extract and plot max daily withdrawals for Durham (PWSID='03-32-010') for each month in 2015

```
#7
durham_data_2015 <- scrape_data_function('03-32-010', 2015)
ggplot(durham_data_2015, aes(x = Date, y = MaxDayUsage)) +
geom_line() +
geom_point() +
labs(title = "Maximum Daily Water Withdrawals for Durham in 2015",
x = "Date",
y = "Max Day Use (MGD)") +
theme_minimal() +
scale_x_date(date_breaks = "1 month", date_labels = "%b")
```



8. Use the function above to extract data for Asheville (PWSID = 01-11-010) in 2015. Combine this data with the Durham data collected above and create a plot that compares Asheville's to Durham's water withdrawals.

```r
#8
asheville_data_2015 <- scrape_data_function('01-11-010', 2015)
combined_data <- bind_rows(asheville_data_2015, durham_data_2015)
ggplot(combined_data, aes(x = Date, y = MaxDayUsage, color = WaterSystemName)) +
geom_line() +
geom_point() +
labs(title = "Comparison of Monthly Maximum Daily Water Usage in 2015",
subtitle = "Asheville vs. Durham",x = "Date",
y = "Max Day Usage (MGD)") +
theme_minimal() +
scale_color_manual(values = c("Asheville" = "blue", "Durham" = "red")) +
scale_x_date(date_breaks = "2 months", date_labels = "%b-%y")
```

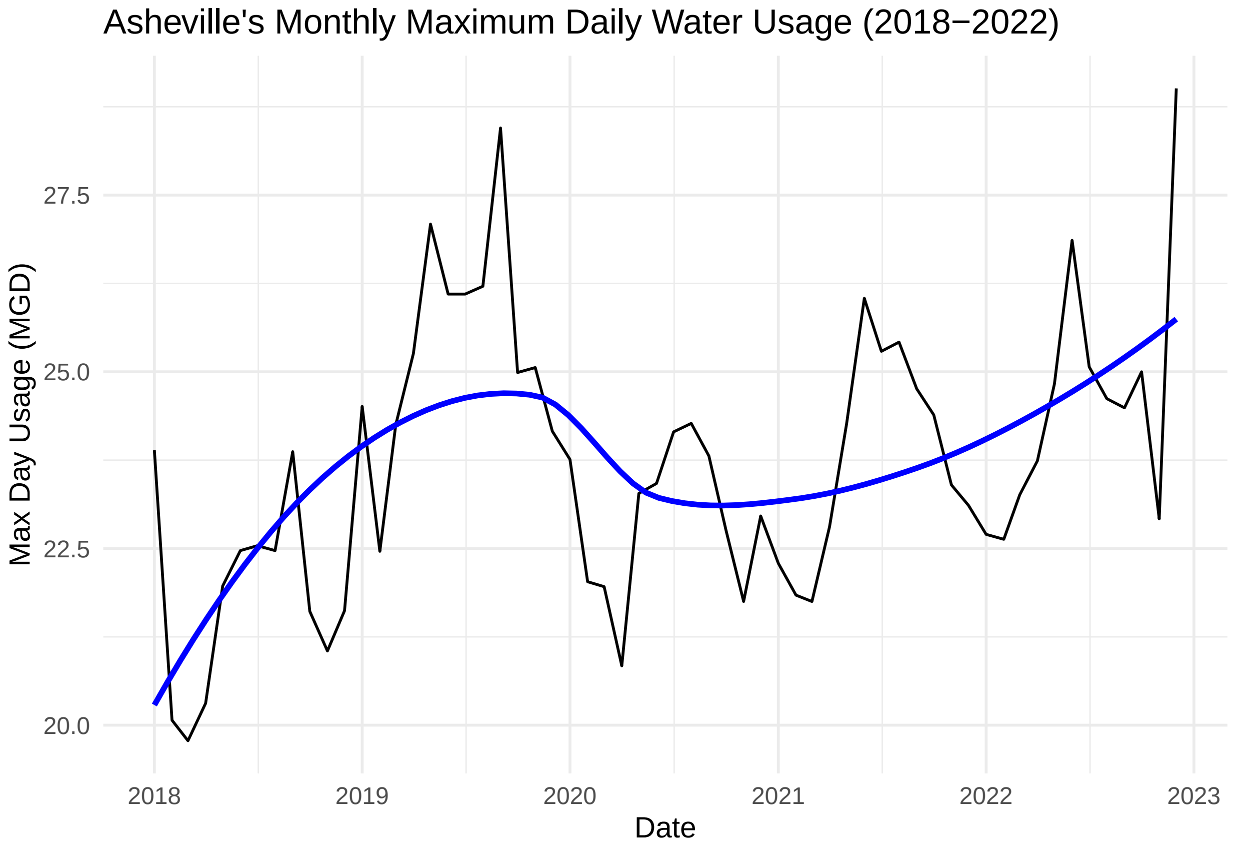


9. Use the code & function you created above to plot Asheville's max daily withdrawal by months for the years 2018 thru 2022.Add a smoothed line to the plot (method = 'loess').

   TIP: See Section 3.2 in the "10_Data_Scraping.Rmd" where we apply "map2()" to iteratively run a function over two inputs. Pipe the output of the map2() function to `bindrows()` to combine the dataframes into a single one.

```
#9
years <- 2018:2022
asheville_data_list <- map(years, ~ scrape_data_function('01-11-010', .x))
asheville_data <- bind_rows(asheville_data_list)
ggplot(asheville_data, aes(x = Date, y = MaxDayUsage)) +
geom_line() +
geom_smooth(method = 'loess', se = FALSE, color = "blue") + # Add smoothed line
labs(title = "Asheville's Monthly Maximum Daily Water Usage (2018-2022)",
x = "Date",
y = "Max Day Usage (MGD)") +
theme_minimal() +
scale_x_date(date_breaks = "1 year", date_labels = "%Y")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Question: Just by looking at the plot (i.e. not running statistics), does Asheville have a trend in water usage over time? > Answer: Yes, it looks that Asheville's water usage has a upward trend over the period from 2018 till the last quarter of 2019. Then, it has downward between 2020 and 2021. Then again upward after that.